Evaluating the C-section Rate of Different Physician Practices: Using Machine Learning to Model Standard Practice

Rich Caruana[1], Radu S. Niculescu[2], R. Bharat Rao[3], Cynthia Simms MD[4]

caruana@cs.cornell.edu; stefann@cs.cmu.edu; bharat.rao@siemens.com; csims@mail.magee.edu

[1]Cornell University, Computer Science, 4157 Upson Hall, Ithaca, NY 14853

[2]Carnegie Mellon University, Computer Science, 5000 Forbes Avenue, Pittsburgh, PA 15213

[3]Siemens Medical Solutions, Inc., 51 Valley Stream Parkway, Malvern PA 19355

[4]Department of Obstetrics, Gynecology, and Reproductive Sciences, University of Pittsburgh, Magee-Womens Hospital, 300 Halket St., Pittsburgh PA 15213

ABSTRACT

The C-section rate of a population of 22,175 expectant mothers is 16.8%; yet the 17 physician groups that serve this population have vastly different group C-section rates, ranging from 13% to 23%. Our goal is to determine retrospectively if the variations in the observed rates can be attributed to variations in the intrinsic risk of the patient sub-populations (i.e. some groups contain more ``high-risk C-section'' patients), or differences in physician practice (i.e. some groups do more C-sections). We apply machine learning to this problem by training models to predict standard practice from retrospective data. We then use the models of standard practice to evaluate the C-section rate of each physician practice. Our results indicate that although there is variation in intrinsic risk among the groups, there is also much variation in physician practice.

1. INTRODUCTION

Our goal is to determine if *groups* of patients seen by different physician practices have different intrinsic risks for C-section. Our approach is as follows: we train a model to predict standard practice using machine learning (in this study, bagged probabilistic decision trees). We use the model to estimate the intrinsic risk of each group by averaging the C-section risk the model predicts for each patient in that group. Differences between the observed and predicted C-section rates indicate physician groups with behavior different from that predicted by the standard practice model.

Intrinsic factors are factors related to patient health that should be used to make care decisions. Our data includes 82 intrinsic factors: pre-pregnancy health-and-physical factors such as maternal age, weight, smoking, diabetes, and prior pregnancy; mid-pregnancy factors such as changes in maternal blood sugar and estimated fetal weight; and labor factors such as maternal blood pressure and fetal distress. These intrinsic factors are the inputs to the model trained to predict C-section. Extrinsic factors are all factors not entailed by these inputs. Extrinsic factors include type of physician practice, type of patient insurance, and patient socio-economic status. The model trained to predict standard practice is allowed to use intrinsic variables to predict patient risk. If the model is accurate, it will compensate for differences between patients (or groups of patients) caused by the intrinsic variables, but will not compensate for differences due to extrinsic variables it did not have access to. This will allow us to determine if the variations in observed C-section rates can be attributed to variations in *intrinsic risk* of the patient sub-populations (i.e., some groups see more ``high-risk C-section'' patients), or if they are due to differences in physician practice (i.e., some groups do C-sections more often).

Section 2 discusses the problem of C-section rate. Section 3 describes our methodology. We use bagged decision trees to train a model of standard practice. Section 4 uses this model to predict the intrinsic risk of different groups of patients. Differences between observed and predicted risk represent a possible difference between physician behavior and standard practice. Section 5 discusses the assumptions made by this approach.

2. BACKGROUND

2.1 Problem Definition

In the U.S. about 17% of births are by C-section. In Europe, the C-section rate is substantially lower, but outcomes do not appear to be worse. Poma notes that the C-section rate in the U.S. increased significantly, yet there has not been a related improvement in neonatal outcomes, suggesting the rate is unnecessarily high [4]. The Pennsylvania Health Care Cost Containment Council notes that cesarean deliveries carry increased risk of complications and longer patient recovery times as well as higher health care costs [3]. The average cost of a C-section in Southwestern PA in 1998 was $7,885 and the average cost for a vaginal delivery was $4,787.

There are medical and financial benefits to a lower C-section rate if outcomes are not adversely affected.

Insurance companies in the U.S. have begun applying financial pressure to lower the C-section rate. One such policy is to pay for a fixed percentage of C-sections. If a practice has a rate higher than the quota, it must make up the difference. If the rate is lower, it makes more profit.

There are problems with using financial pressure to reduce C-sections. One problem is the tragedy of the commons: individual doctors often have incentives not to lower their C-section rate, even though *groups* of physicians would benefit by lowering their group rate. This problem is complicated by the fact that doctors do not see patients of equal risk. Some doctors specialize in high-risk pregnancies and thus should have a higher C-section rate. To evaluate practices fairly, an objective model needs to be developed that can predict whether or not patients *should have* received C-section.

In [1], the C-section rates of different hospitals are compared after correcting for the fact that hospitals saw patients with different risks. They constructed a logistic regression model to predict patient risk. Recent studies by members of our group indicated that machine learning methods such as decision trees and neural nets might be preferable to logistic regression [2].

Commonly agreed upon C-section risk factors were used in [3] to distinguish between high and low-risk patients. In [4], an attempt was made to determine obstetrician characteristics that affect C-section rate. The extrinsic factors correlated with lower C-section rates were: younger obstetrician age, graduation from a domestic medical school, belonging to a group practice, and a smaller number of births.

2.2 Magee C-section Database

The database we use is from Magee-Women's Hospital. It contains 22175 patients from 1995-1997. Each record has 144 attributes, of which we use 82 as intrinsic inputs for learning. Each patient in the database is from one of 17 different physician group practices. The goal of our work is to identify physician groups for which the actual C-section rate and the rate predicted by our model of standard practice differ significantly.

3. APPROACH

In preliminary experiments we tried several different types of decision trees and neural networks. We found that the MML decision trees in Buntine's IND software performed particularly well [5]. MML decision trees are grown to full size (often many thousands of nodes) and are not pruned. Instead, Bayesian smoothing is applied to the tree to yield predictions at leaf nodes that are a function of the class probabilities along the entire path leading to each leaf node. We often find that MML trees excel at predicting probabilities. To further improve the predicted probabilities, we applied bagging [9],[10] to the MML decision trees. See [9] for a description of why bagging usually improves the quality of probabilities predicted by decision trees. The bagged trees were trained as follows:

1. Bootstrap samples are drawn to form 100 train sets $T_1 \ldots T_{100}$.
2. An MML decision tree is grown on each T_i.
3. For each example in the dataset, we average the predictions of the trees that *did not contain* this example in their training set.

4. RESULTS

It is critical that the probabilities generated by the model trained to predict standard practice are well calibrated. Suppose the model of standard practice was excellent at ordering patients by relative risk (and thus had excellent ROC[1] performance), but the probabilities it predicted were consistently low (high). Then the aggregate risk obtained by averaging the predicted probabilities for a group of patients would consistently underestimate (overestimate) the true aggregate risk and most physician practices would appear to have C-section rates higher than (lower than) is warranted by the patients they see. But this is not a real problem because it is easy to force the average predicted rate to equal the average observed rate by normalizing.

A more serious concern is that the model of standard practice must be well calibrated in the low and high-risk tails of the population. For example, suppose the model probabilities are normalized, but that the model tends to predict somewhat low probabilities for high risk patients and somewhat high probabilities for low risk patients. The model might predict p=0.6 for patients that have true risk of C-section p=0.8. This model might be accurate, and have good ROC, but when applied to a group of patients with disproportionately high risk, would underestimate the group's aggregate risk, thereby causing us to suspect that the group was performing C-sections at an unwarranted high rate.

Poor calibration in the tails is common. Before using the bagged decision tree model of standard practice, we must

[1] The Receiver Operator Characteristic (ROC) curve is a plot of the true-positive rate vs. the false-positive rate as the prediction threshold is varied from 0 to 1. The area under the ROC curve (AUC) is a statistic that commonly is used to summarize the performance of a model. AUCs closer to 1 indicate that the model is better at predicting higher risk for patients that truly have elevated risk.

verify that it has good calibration. To do this, the risk interval [0,1] was split in 19 overlapping subintervals of width 0.1: [0,0.1], [0.05,0.15],…,[0.9,1]. The patients with predicted risk falling in each subinterval were placed in each subinterval and used to calculate the average observed C-section rate for that subinterval. Figure 1 shows a plot of the observed C-section rate for each subinterval plotted against the predicted C-section rate for that subinterval. The plot is remarkably true to the 45 degree line, indicating excellent calibration. The average absolute difference between the predicted risk and observed C-section rate is an extremely low 0.013.

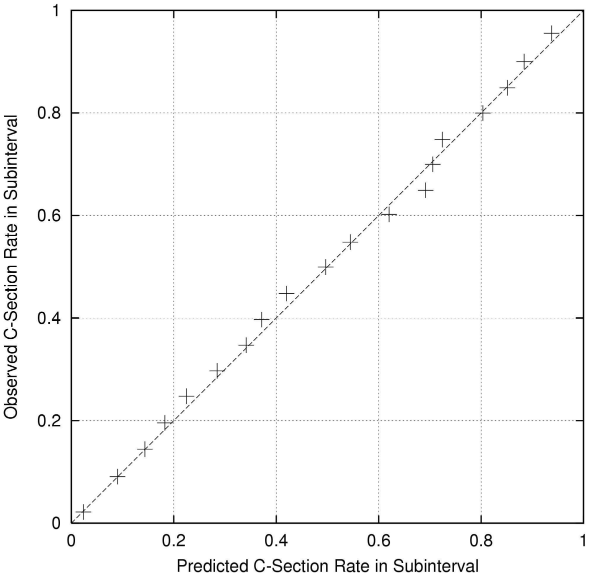

Figure 1. Calibration of the Standard Practice Model

To verify that the model we train is good at predicting the standard practice, we also measured its accuracy and ROC on test cases held out of the training sets. The accuracy of the model on these test cases is 87%. The ROC Area of the model on the same test cases is 0.9233. These figures suggest the model is very good at predicting standard practice.

After checking the model's calibration, we used the model to predict the aggregate risk of each of the 17 physician practices by averaging the predicted risk of all patients in each practice. This yields the expected C-section rate for each group, corrected for the risk of the patients in that group. Figure 2 is a scatter plot of the observed C-section rate vs. the predicted rate for each of the 17 physician groups. Points that fall near the diagonal have an observed C-section rate similar to that predicted by the models. Physician groups having lower C-section rates than the models predict fall in the upper left. Physician groups having C-section rates higher than the models predict fall in the lower right.

Most physician groups fall near the diagonal, indicating that their C-section rates are comparable to the rates predicted by the standard practice model. Physician groups H, J, and O, however, exhibit high C-section rates that may not be warranted. O's high rate appears to be somewhat justified because the model predicts the patients in group O have the highest risk of all 17 groups. Groups H and J, however, appear to consist of patients with lower than average risk, yet their C-section rate is well above average. Interestingly, physician group G exhibits a surprisingly low C-section rate given the predicted aggregate risk of its patient population.

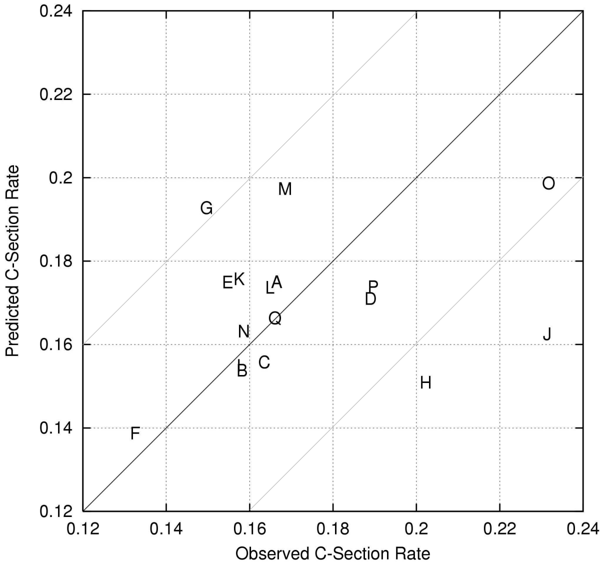

Figure 2. Scatter Plot of the Predicted C-Section Rate vs. the Observed Rate in each Physician Practice

Figure 3 shows a scatter plot of the ROC Area of the standard practice model evaluated individually on each physician group plotted against each group's observed C-section rate. The AUC for group H is lower than that of the other groups. Either the model of standard practice makes less accurate predictions for patients in group H despite the model's excellent calibration, or physicians in group H make decisions about C-section somewhat differently than is the practice in the other groups.

5. DISCUSSION

5.1 Assumptions

Our approach makes several assumptions. One is that by giving models intrinsic variables as inputs, they will be accurate enough to compensate for these factors, yet unable to compensate for extrinsic factors not given as inputs. As with most machine learning models, the difference between the observed and predicted risk is attributed to the risk due to these extrinsic factors. These assumptions are not fully justified because we may not capture all variables that relate to the health of the patient (missing important inputs) and because some of the extrinsic factors may correlate with the intrinsic variables (possibly allowing the model to partially account for extrinsic factors).

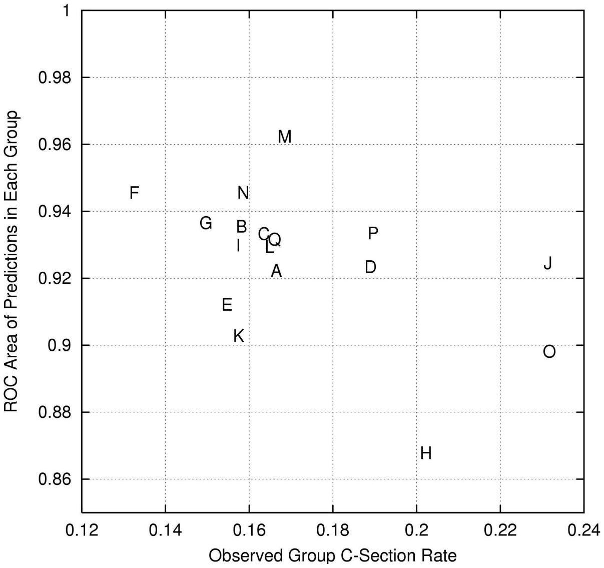

Figure 3. Scatter Plot of the Predictions ROC Area vs. Observed C-Section Rate for each Physician Practice

5.2 Predicting Care vs. Standard Practice

There are interesting differences between learning intended to make predictions for individual patients, and learning models of standard practice to retrospectively assess aggregate risk as done here. One difference is that when making predictions for individual patients, overfitting must be avoided because it increases variance more than it reduces bias [9], thus hurting generalization performance. It is very important not to make mistakes when making predictions that will affect the care of individual patients. When using learning to retrospectively assess aggregate risk, however, this tradeoff is somewhat different. Because the aggregate risk of a population of patients averages model predictions over that population, variance is reduced by the average and is thus less of a concern. Some overfitting can be tolerated if it will reduce bias and improve model calibration. Calibration in the low and high-risk tails of the distribution is particularly important. We use bagging not because it is effective at reducing variance, but because experience suggests that it significantly improves the calibration of decision trees. It would be interesting to extend the usual bias-variance decomposition so that the tradeoff between bias and variance can be better optimized for making aggregate predictions.

6. FUTURE WORK

6.1 Standard Practice vs. Best Practice

In this paper we use machine learning to evaluate how the decisions made by different physician groups compare to the standard practice of peer physicians. We do not examine the quality of health care that results from the standard practice. An important extension of this work is to compare outcomes in the different physician groups to determine if differences in C-section rate correlate with quality of care. Specifically, it would be informative to see if the difference between observed C-section rate and predicted rate correlates with quality of care.

The ultimate goal of our work is to provide an evidence-based means of reducing C-section rate in order to lower health care cost and improve maternal outcomes without worsening fetal outcomes. The best evidence we know of that C-section rate can be lowered without adversely affecting outcomes comes from other countries that have lower C-section rates, but comparable outcomes. Access to a database of patients from other countries might allow us to perform analyses not possible with the U.S. database alone.

One limitation of our current method is that it does not permit us to estimate confidence intervals for the aggregate risks predicted by the models. Although all of our experiments yield results consistent with those presented here, it is important for us to develop a procedure for estimating the reliability of standard practice models.

6.2 Other Applications

The approach of training well calibrated models to predict standard practice and then using these models to assess the aggregate risks of different subpopulations is applicable to other problems in medical decision making. For example, we might determine if different subpopulations of patients with heart disease receive different rates of coronary bypass because they have different risk, or because of other factors such as patient socio-economic group, care provider (e.g., small practice vs. large practice, or specialty practice vs. general practice), or health care insurance (e.g., HMO vs. PPO (pay-per-use)).

7. SUMMARY

We use decision trees with Bayesian smoothing and bagging to train models of *standard practice* for C-section. We use the models of standard practice to perform a retrospective evaluation of the C-section rate of different physician practices. Our goal is not to make accurate predictions for single patients, but to make accurate *aggregate* predictions for *groups* of patients. (By "*accurate*" we mean in accordance with common practice, not necessarily medically correct.) Because we are interested in accurate aggregate predictions, it is important that our models be well calibrated. We find that bagged decision trees yield excellent calibration in this domain. Because the calibration is so good, we believe the resulting models are not biased for or against any one group or type of patients.

Using the models to estimate the aggregate risk for the 17 different physician practices yields interesting results. Our analysis suggests that several practices who had a C-section rate 3-6% higher than the population average probably do not have a patient population with enough elevated risk to warrant this C-section rate. In fact, one of these practices sees patients whose aggregate risk for C-section appears to be *below* the average risk. Other practices seeing these same patients probably would do fewer C-sections. At least one of the groups with elevated C-section rate, however, has a patient population that partially justifies the high C-section rate. Their patient population truly is at elevated risk and warrants a higher C-section rate.

Other factors not included in the trained models such as patient and physician preferences, or the type of health care funding, might explain why some groups receive more C-sections. Most patient groups have predicted C-section rates similar to the observed rates, suggesting that most physician groups are performing C-sections at a rate in accordance with standard practice. Interestingly, there is one group that had 4% *fewer* C-sections than the model of standard practice predicts might be warranted. If this lower C-section rate does not increase the adverse outcomes, this practice may provide insight on how to safely lower the C-section rate.

Acknowledgement. Rich Caruana received partial support from the Intelligent Information Systems Institute (IISI). Radu Stefan Niculescu was supported by a Graduate Fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation and by National Science Foundation (NSF) grant no. CCR-0122581. Matt Troup and Peng Jia helped perform preliminary experiments on this problem.

REFERENCES

[1] Bailit JL, Dooley SL, Peaceman AN. Risk Adjustment for Interhospital Comparison of Primary Cesarean Rates. J. Obstetrics and Gynecology 1999; 93:1025-1030.

[2] Sims CJ, Myen L, Caruana R, Rao RB, Mitchell T, Krohn M. Predicting cesarean delivery with decision tree models. J. Obstetrics and Gynecology 2000;183:1198-1206.

[3] Pennsylvania Health Care Cost Containment Council. C-section and Vaginal Deliveries In Southwestern Pennsylvania. Report dated July 1999.

[4] Poma PA. Effects of obstetrician characteristics on cesarean delivery rates: A community hospital experience. J. Obstetrics and Gynecology 1999; 180:1364-1372.

[5] Provost F, Fawcett T, Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. Proceedings of the Fifteenth International Conference on Machine Learning 1998.

[6] Buntine W, Caruana R. *Introduction to IND and recursive partitioning*. Technical Report FIA-91-28, RIACS and NASA Ames Research Center, Moffett Field, CA, 1991**.**

[7] Caruana R. An Non-Parametric EM-Style Algorithm for Imputing Missing Values. Proceedings of Artificial Intelligence and Statistics 2001.

[8] Zell A, Mache N, Hubner R et al. "SNNS:Stuttgart Neural Network Simulator," Tech. Rep. 3/93, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Fed. Rep. of Germany, 1993.

[9] Bauer E, Kohavi R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 1999;36:105-139.

[10] Breiman L. Bagging Predictors. Machine Learning 1996;24:123-140.